DOCUMENT RESUME

ED 211 083 IR 009 905

AUTHOR	Katzer, Jeffrey; And Others
TITLE	A Study of the Impact of Representations in Information Retrieval Systems. Annual Technical Report.
INSTITUTION	Syracuse Univ., N.Y. School of Information Studies.
SPONS AGENCY	National Science Foundation, Washington, D.C. Div. of Information Science and Technology.
PUB DATE	May 81
GRANT	IST-79-21468
NOTE	67p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Analysis of Variance; Databases; *Information Retrieval; Models; *Online Systems; *Performance Factors; *Relevance (Information Retrieval); Research Methodology
IDENTIFIERS	INSPEC

ABSTRACT

This study conducted to determine representation impact on information items retrieval in terms of precision and recall performance and overlap used the INSPEC "Computers and Control Abstracts" loaded on DIATOM, an online retrieval system based on DIALOG, as the database to be searched. Sixty-nine users provided 84 queries which were searched for high recall by intermediaries under each of seven representations: title only, abstract only, descriptors, identifiers, title and abstract, stemmed title and abstract, and the descriptor and identifier fields. Copies of the retrieved citations and abstracts were sent to users for judging relevance. Then the seven representations were tested using a latin square design on the 84 queries. Measures of recall, precision, and total retrieval of citations were analyzed using standard analysis of variance computations; the performance measures and overlaps findings are presented in detail. The results confirm earlier observations that there is relatively little difference in performance among the representations and relatively little overlap. Plans for observations and findings replication of the first phase and theory development for Phase II are described. Eleven tables, 19 references, and five appendices are provided. (RBF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED211083

A STUDY OF THE

IMPACT OF REPRESENTATIONS

IN INFORMATION RETRIEVAL SYSTEMS

Annual Technical Report

May 1981

This material is based on research supported by the National Science Foundation, Division of Information Science and Technology, under Grant IST 79-21468. The opinions, findings and conclusions or recommendations expressed in this report are those of the authors and do notnecessarily reflect the views of the National Science Foundation

School of Information Studies
Syracuse University
Syracuse, New York 13210

IR009905

This report was written by

Jeffrey Katzer, Michael McGill,
Judith A. Tessier, William Frakes
and Padmini DasGupta

PROJECT STAFF

Co-Principal Investigators	Jeffrey Katzer Michael McGill
Faculty Associate	Judith A. Tessier
Graduate Associate	William B. Frakes
Graduate Assistants	Padmini DasGupta Cheryl McAfee
Project Secretary	Margaret Montgomery
Consultants, Phase I	Terry Noreault Matthew Koll Robert Waldstein

ABSTRACT

A key element of an information system is the representation of the information items. Studies have found that, when using precision and recall performance measures, the differences among various representations are not critical. Evidence does indicate that the actual items retrieved vary significantly from representation to representation. This study will determine the impact of representation on the retrieval of information items in terms of performance and overlap and suggest performance limits for an information system, given a specific representation.

This interim report describes Phase I of the project. Seven representations were tested using a latin aquare design on 84 queries. The INSPEC Computers and Control Abstracts was the study data base loaded on the DIATOM system. The data generally confirm the earlier observed data: overlaps were again small. Plans for replication and theory development in Phase II are described.

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
II.	OBJECTIVES	5
III.	RETRIEVAL ENVIRONMENT	5
	A. Data Base	5
	B. Retrieval System	6
	C. Search Intermediaries	7
	D. Users and Queries	7
	E. Relevance Judgments	9
IV.	METHODOLOGY	10
	A. Variables	10
	B. Procedure	13
	C. Design and Analysis	14
V.	RESULTS	16
	A. Analysis of Performance	16
	B. Analysis of Overlaps	21
VI.	PHASE II PLANS	33
	REFERENCES	36
	APPENDICES	40
	A. Training Materials	
	B. Form for Relevance Judgments	
	C. Directions to Users	
	D. Forms for Searcher, Attached to Query	
	E. Latin Square Design	
	F. AOV Summary Results	

TABLES

		Page
1.	Characteristics of Users	8
2.	Performance and Overlap Comparisons	17
3.	Means and Standard Deviations by Representations	19
4.	Mean Performed by Representation Across Queries	20
5.	Symmetric Pairwise Overlaps	23
6.	Asymmetric Pairwise Overlaps	24
7.	Union Pairwise Overlaps	25
8.	Representations Ordered by Incremental Improvement	27
9.	Representations Ordered by Incremental Improvement *	28
10.	Recalls and Unique Contributions of 7 Representations	30
11.	Unique Contributions of 4 Representations *	31

I. INTRODUCTION

This report presents the interim results of the Document Representation study. The report will describe the research background and objectives, procedures used during the first phase of the study, results of the first phase, and plans for the second phase. The document representation study is designed to provide fundamental knowledge of the effect of the representation of information items on information system performance.

Past studies have found that, when using precision and recall performance measures, the differences among various representations is not critical. Studies to date have examined the precision and recall performance of two or more representations. The unifying element of these studies is a search for a "better" representation. That is, given a specified environment and using a particular set of queries, which representation performs better in terms of precision and recall? In these studies, no one representation clearly outperforms others. But studies have shown that when using a particular representation it is possible to employ techniques to enhance the performance of that representation.

This study takes as its departure evidence that performance measures have masked real and systematic differences among the representations. Specifically, different representations result in the retrieval of different items. Two previous studies support the hypothesis.

The Ranking Project (MCGILL) examined the specific items retrieved from each of the representations used in that study. The same searcher using different representations for the same information need statement had an overlap of retrieved items totalling 14%. Different searchers using different representations had an overlap of the retrieved set of 5%. That is, this study found that using the free representation or the controlled representation did not affect performance measures, but it did impact the actual items retrieved by the system. The user can expect approximately the same number of relevant documents using either representation - however, the actual documents retrieved are not the same.

SMITH examined the combination of document representation and similarity measure. Her work was conducted using a subset of the INSPEC data base. Using the representation of a document as a query, she examined seven different representations. SMITH did not investigate

performance measures, but did report non-symmetric overlap. Non-symmetric overlap was defined as

$$\frac{n(A \cap B)}{n(B)} \text{ and } \frac{n(A \cap B)}{n(A)}$$

The non-symmetric measure indicates the direction of the overlap. Nonsymmetric overlap measures among the retrieved sets ranged from a mean overlap measure of .489 (or approximately 50% of the documents were in sets retrieved by both representations) to a mean of .004 (or only 0.4% of the documents were retrieved by both representations.

These studies indicate the potential importance of the selection of representations of information items. However, neither of the above studies is conclusive or generalizable. This study is designed to build on the previous findings and to ultimately develop a theoretical model accounting for representation differences.

II. OBJECTIVES

The assessment of the various representations is concerned with a number of specific objectives:

(1) To determine if the information items retrieved by the differing representations are significantly and substantially different.

(2) To assess the effectiveness of representations or combinations of representations.

(3) To develop and test a theoretic model sufficient to explain any differences in information retrieval system operation based on changes in the representation of information items.

At the conclusion of the study, an information scientist should be able to discern the relative impact of a particular representation. The data should indicate which representations are redundant or may be used in place of another, and which representations may be used in combination to enhance a particular aspect of system performance, such as recall. Finally, it may be possible to specify upper bounds of particular performance measures given a particular representation.

III. RETRIEVAL ENVIRONMENT

A. Data Base

Permission was granted by the Institution of Electrical Engineers to use the Computer and Control Abstracts portion of the INSPEC data base. Altogether 12,000 documents formed the data base used in this study. These constituted the September - December 1979 issues of Computer and Control Abstracts. The choice of this data base and its size provided enough topic specificity to ensure that a reasonable number of documents would be retrieved in each representation.

Each document consisted of a series of bibliographic citation fields, an abstract, and some indexing information. The format of each document record as it was printed upon retrieval is as follows:

DNnumber (abstract numbers from INSPEC journals)
Title
Authors (separated by commas)
Source field: as follows
Publication: (volume and issue number) (part number) pagination data
Following this may be information in [ ]. This is information on the cover-to-cover translation as follows: [publication; (volume and issue) pages, date] (type of unconventional media) (availability) (Title of conference) location of conference) (sponsoring organization) (date) language.
Abstract
Indexing information

B. Retrieval System

DIATOM, an on-line retrieval system which was designed to simulate most of the features of Dialog, was used to conduct all the searches in this study. DIATOM was designed and programmed by Bob Waldstein, a PhD student at the School of Information Studies.

The major differences between DIATOM and those of DIALOG are listed below.

1. Diatom permitted the searchers to log on directly to a particular representation. All search statements were subsequently restricted to that representation only.

2. The system included a stemmer used for the stem representation.

3. To restrict a search to a particular language, a Limit /ENG (for English) was used.

4. Adjacency (nW) could not be used with either truncation or stemming.

5. Adjacency at times ran very slow; the field operator (F) could be used instead.

C. Search Intermediaries

A total of seven intermediaries were required for the research design. All of the intermediaries used in the study were professional librarians or information brokers with experience using computerized retrieval systems; all had had some experience using DIALOG.

All intermediaries took part in a one day long training session. Afterwards, each intermediary was required to familiarize himself with the system and make at least 14 searches to the data base. A copy of the training materials furnished the intermediaries is provided in Appendix A.

D. Users and Queries

Originally the study specified 98 users, each of whom was to provide a single interest statement or query. However, because of difficulty in obtaining users, the study was reduced to 84 queries. Users were solicited from the Syracuse University community and institutions concerned with information retrieval. Table 1 indicates characteristics of the users. Our objective in accepting users was to come as close as possible to criteria used in operational search services so that queries and relevance judgments could plausibly be generalized.

TABLE 1
Characteristics of Users

Affiliation	No. of Users	Faculty	Students	Sci/Eng	Others	No. of Queries
Syracuse U.	35	26	8	0	1	41
General Electric	1	0	0	1	0	4
Univ. of Illinois	5	2	3	0	0	5
Univ. of Louisville	9	0	0	0	9	14
National Bureau of Standards	6	0	0	6	0	6
OCLC,INC.	5	0	0	5	0	6
Environmental Protection Agency	6	0	0	6	0	6
OTISCA Industries	1	0	0	0	1	1
SUNY College of Environ. Sciences & Forestry	1	0	1	0	0	1
	69	28	12	18	11	84

*Altogether, 69 individuals served as users in this study. 11 of these individuals submitted more than one query: 8 users submitted 2 queries, 2 users submitted 3 queries and 1 user submitted 4 queries.

E. Relevance Judgments

Relevance judgments were obtained from the users for all documents retrieved for the query.* A four point scale was used with "1" and "2" indicating relevant, "3" and "4" indicating non-relevant. The instructions which accompanied the search results are provided in Appendix B.

*After repeated attempts, four users did not return their relevance judgments. In these few cases we identified other individuals who presumably could make relevance judgments in the specific topic area of the query. These surrogate users made the relevance judgments.

IV. METHODOLOGY

A. Variables

The key experimental or independent variable was the representation used in searching the data base. Seven representations were chosen:

TT - terms in title only.

AA - terms in abstract only.

DD - descriptor terms only.

II - identifier terms only.

TA - terms in title and abstract only.

ST - stemmed terms in title and abstract only.
(The computer automatically takes the logical root of any entered term.)

DI - terms in descriptor and identifier fields.

The major dependent variables were performance measures (recall and precision) and measures of overlap. In addition, a count of the total number of retrieved documents was also analyzed. A more precise description of each of the measures is given below.

RECALL. The recall ratios were formed by dividing the number of relevant documents retrieved by each

representation by the total number of relevant documents retrieved by all seven representations. Two versions of recall were computed.

Recall-1: defined a relevant document stringently. The user had to judge the document to be "most relevant" -- that is, rate it a "1" on the four point scale.

Recall-2: defined a relevant document more broadly. The user could rate it either as a "1" or a "2" on the four point scale.

PRECISION. The precision ratio was formed by dividing the number of relevant documents retrieved by each representation by the total number of documents retrieved by that representation. Two versions of precision were computed.

Precision-1: defined a relevant document stringently-- a "1" on the four point scale.

Precision-2: defined a relevant document more broadly -- a "1" or a "2" on the four point scale.

TOTAL-RETRIEVED. This measure is simply the total number of documents retrieved by each representation; it is the denominator of the precision ratio. It was included because it is an indication of user effort required to read the output from the system.

SYMMETRIC-OVERLAP. For two representations, A and B, this measure is computed by dividing the number of documents retrieved in common by both representations by the total number of documents retrieved by both representations. Or more formally, it is the number of retrieved documents in the intersection of the two representations divided by the number of retrieved documents in the union of the two representations. Three versions of the symmetric-overlap were computed.

Symmetric-1: counted only highly (i.e. "1" on the four point scale) relevant documents retrieved.

Symmetric-2: counted all (i.e. "1" or "2") relevant documents retrieved.

Symmetric-all: counted all documents retrieved.

ASYMMETRIC-OVERLAP. For two repesentations, A and B, this measure is computed by dividing the number of documents retrieved by both representations by the number of documents retrieved by one of the representations. A smaller asymmeric overlap indicates a greater degree of independence of one representation (in the denominator) from the other representation. And, as is the case of the symmetrical measure, there are three versions of this measure: most relevant, all relevant, and all documents.

UNION-OVERLAP. For two representations, A and B, this measure is computed by dividing the number of documents

retrieved by either of the representations by the number of documents retrieved by all seven representations. It is the number of retrieved documents in the union of the two representations divided by the number retrieved in the union of all seven representations. Thus, the union overlap can be viewed as a recall ratio for a combination of representations. This measure extends to more than two representations and three versions of it can be computed: most relevant, all relevant, and all documents retrieved.

B. Procedure

Queries were obtained from users one at a time (see Appendix C for the directions given users). The queries were used as submitted; they were not screened for appropriateness to the data base or for on-line searching. Each of the seven searchers was given a photocopy of the search request. For each query, each searcher received instructions which specified the one representation that searcher was to use for that query. Representations were assigned to searchers on each query according to the latin square design.

Thus, each of the 84 queries was searched under each of the seven representations; in total, seven searches (each using a separate representation) were carried out for each

of the 84 queries.

Searchers used DIATOM to retrieve documents. Searchers were instructed to carry out a "high-recall" search, retrieving a maximum of fifty documents. The directions given to each intermediary is given in Appendix D.

After all seven intermediaries completed a query, the seven retrieved document sets were merged into a single listing and placed in reverse accession number order. The listing consisted of the citations and abstracts of all retrieved documents. No clue was present which indicated either the searcher or the representation.

Two copies of this listing were produced. Both copies were sent to the user with instructions (see Appendix B) to make relevance judgments on one copy and return that copy to the project. The second copy was for the user.

C. Design and Analysis

The overall design can be characterized as a 7x7 latin square replicated 12 times. The full design is given in Appendix E.

The measures of recall, precision, and total-retrieved are analyzed using standard analysis of variance computations. The design and the analysis control for extraneous variables and can identify separate effects for representations, intermediaries, and if desired,

replications. Approximately ten percent (66) of the precision results had to be excluded from the analysis because no documents were retrieved for a given query under a given representation. Fourteen queries had to be excluded from all Recall-1 analyses, and seven from the Recall-2 analysis, because in each situation no relevant documents were retrieved.

The overlap measures may have been adversely affected by the latin square design. Because each pair of representations for a given query were searched by different intermediaries, there is a possibility that the overlap measures confound representations with intermediaries. Keeping this concern in mind, we will compute and interpret the results of the overlap analyses. The overall design will be changed for the second phase of this study in order to prevent this possibility.

V. RESULTS

Our initial concern was to determine if the results from this study repeated the pattern noted earlier: relatively little difference in performance among the representations coupled with relatively little overlap. Table 2 presents these results. It is apparent that these results do repeat the pattern observed in other studies. Though some performance measures are significantly different, none of the differences exceed 18% -- which is clearly within the range of values reported in the literature. The overlaps range from a low of about 6% to a high of about 17%; these also correspond to the earlier results.

The remaining part of this section presents these findings in more detail. First the performance measures will be considered. Then the study of overlaps will be presented.

A. Analysis of Performance

Descriptive summary statistics for the five performance measures are presented in Table 3. The means were tested for statistically significant differences (see Appendix F for the AOV Summary Tables). Representations differed significantly in the Recall-1, Recall-2, and Total-Retrieved scores. The bottom of Table 3 indicates that descriptors (DD) and titles (TT) perform rather poorly as

TABLE 2
Performance and Overlap Comparisons
Between the "Best" and the "Worst" Representations

	REC-1	REC-2	PRE-1	PRE-2	TOT-RET
"Best" Rep.	.404	.321	.264	.422	19.833
"Worst" Rep.	.229	.200	.173	.336	12.429
Difference	.175*	.121*	.091	.086	7.404*
Symmetric overlap**	.155	.138	.172	.150	.057

*Difference is statistically significant at .05 level

**Symmetric overlap figures are taken from TABLE 5 using the pairwise overlap between the "Best" and "Worst" for each performance measure, e.g. the pairwise overlap for Relevant "1's" for TA ("Best") and DD ("Worst") is used for Column 1, REC-1.

representations on the recall measures, while identifiers (II) and title-abstracts (either TA or ST) perform much better.

Even though no pairs of representations differed significantly in either precision measure, it is useful to include some consideration of precision into these findings. Considering all five measures, the descriptor (DD) representation performs uniformly poorly on the recall and precision measures while title-abstract (TA) performs reasonably well on them -- though not as strongly as DD's negative performance. Interestingly, the free text words assigned by indexers (II) perform moderately well over all five measures. Stemming (ST) which would tend to increase the total number retrieved performs quite well on the recall measures, but poorly on the precision measures. The title representation (TT) shows the opposite pattern -- high on the precision measures (and Tot-Ret) and low for recall. The other representations fluctuate quite a bit over the five measures.

The recall and precision means given in Table 3 are the average of individual ratios -- each query contributed equally to the final average. Another way to compute the average performance values is to compute the ratio last. For example, for Recall-1, sum the number of relevant documents retrieved from all 70 queries using a particular representation and divide this total by the number of

TABLE 3
Means and Standard Deviations by Representations**

Representation	REC-1	REC-2	PRE-1	PRE-2	TOT-RET
DD (descriptor)	0.229 (70) .319	0.200 (77) .257	0.123 (62) .260	0.336 (62) .330	13.238 (84) 15.824
AA (abstract)	0.365 (70) .314	0.270 (77) .241	0.197 (77) .255	0.352 (77) .315	17.488 (84) 16.850
TA (title and abstract)	0.404 (70) .317	0.290 (77) .236	0.224 (78) .286	0.352 (78) .318	18.583 (84) 16.245
DI (descriptor and identifer)	0.330 (70) .328	0.284 (77) .284	0.221 (75) .270	0.361 (75) .300	16.369 (84) 16.166
ST (stemmed title and abstract)	0.392 (70) .352	0.317 (77) .263	0.188 (81) .231	0.338 (81) .291	19.833 (84) 15.814
TT (title)	0.273 (70) .292	0.205 (77) .207	0.264 (70) .335	0.422 (70) .370	12.429 (84) 13.744
II (identifier)	0.339 (70) .323	0.321 (77) .276	0.218 (79) .282	0.403 (79) .334	16.131 (84) 15.181
Minimum difference between means that are significantly different at .05.*	0.133	0.106	-----	-----	5.450
Pairs of representations that differ	DD<TA DD<ST DD<AA	DD<II DD<ST TT<II TT<ST	none	none	DD<ST TT<ST TT<TA

*Using Tukey's HSD procedure. See Appendix F for details.

**The three values given in each cell of the table are respectively the mean, the sample size, and the standard variation.

TABLE 4
Mean Performance by Representation
Across Queries

Representation	REC-1	REC-2	PRE-1	PRE-2
DD (descriptor)	0.207	0.216	0.173	0.335
AA (abstract)	0.328	0.283	0.181	0.332
TA (title & abst)	0.369	0.294	0.192	0.324
DI (descr & ident)	0.309	0.268	0.182	0.336
ST (stemmed TA)	0.304	0.281	0.148	0.291
TT (title)	0.285	0.229	0.221	0.378
II (identifier)	0.348	0.306	0.208	0.389

representation and divide this total by the number of relevant documents retrieved from all 70 queries using all seven representations. This is a more conservative approach and these values can never exceed the values presented in Table 3. This approach is useful, however, because the unique contribution of single (perhaps atypical) queries is removed. The average values computed in this manner are presented in Table 4. There are several parallels between the patterns in the two tables. Again, the II representation performs well on all four measures. Descriptors (DD) still show an overall poor performance and title-abstract (TA) performs well (though the similarity is weakened in the precision-2 measure). Titles (TT) have the same pattern here as in Table 3, while stemming (ST) is not quite as good in the recall measures and is just as poor in the precision measures.

B. Analysis of Overlaps

The simplest analysis of overlaps is pairwise, comparing each representation with every other representation. Tables 5, 6, and 7 contain the pairwise overlaps for symmetrical, asymmetrical, and union overlap. Each table reports the overlap for relevant documents (only those judged a "1", and those judged a "1" or a "2") and for all documents.

As might be expected, the pairwise overlaps decrease as the number of documents under consideration increases. That is, the average overlap is highest when only most relevant documents are included; it is lowest when all documents are included.

The major finding in these data is that the overlaps are quite small as indicated by the averages. This is true even between representations that should have retrieved very similar sets such as abstract (AA) and title-abstract (TA) or descriptor (DD) and descriptor-identifier (DI). One possible explanation for the size of the overlaps is searcher differences. The analysis of variance tables (see Appendix F) support this contention; they show that between searcher differences accounts for one of the largest portions of the variance. However, the data in the ranking study (MCGILL) cast doubt on the contention that searchers are the sole or major cause of the low amount of overlap. In the ranking study, overlaps between different representations searched by the same searcher only equalled 14% for retrieved documents. That figure certainly falls in the range of values reported here.

Going beyond pairwise overlaps, the question arises as to the optimum combination of representations, or more precisely, the optimum ordering of representations. That

TABLE 5
Symmetric Pairwise Overlaps

	AA	TT	TA	ST	II	DI	DD	AVG
Version - Most Relevant								
AA	1.000	0.181	0.270	0.313	0.212	0.217	0.125	.220
TT	0.181	1.000	0.227	0.178	0.236	0.209	0.172	.200
TA	0.270	0.227	1.000	0.307	0.208	0.236	0.155	.234
ST	0.313	0.178	0.307	1.000	0.179	0.201	0.115	.215
II	0.212	0.236	0.208	0.179	1.000	0.314	0.173	.220
DI	0.217	0.209	0.236	0.201	0.314	1.000	0.270	.241
DD	0.125	0.172	0.155	0.115	0.173	0.270	1.000	.168
Version - All Relevant								
AA	1.000	0.141	0.215	0.235	0.167	0.186	0.112	.176
TT	0.141	1.000	0.154	0.133	0.173	0.172	0.150	.154
TA	0.215	0.154	1.000	0.245	0.167	0.173	0.114	.178
ST	0.235	0.133	0.245	1.000	0.138	0.137	0.081	.161
II	0.167	0.173	0.167	0.138	1.000	0.242	0.138	.171
DI	0.186	0.172	0.173	0.137	0.242	1.000	0.258	.195
DD	0.112	0.150	0.114	0.081	0.138	0.258	1.000	.142
Version - All Documents								
AA	1.000	0.064	0.148	0.138	0.112	0.103	0.046	.102
TT	0.064	1.000	0.072	0.057	0.086	0.080	0.068	.071
TA	0.148	0.072	1.000	0.156	0.096	0.092	0.052	.103
ST	0.138	0.057	0.156	1.000	0.077	0.063	0.033	.087
II	0.112	0.086	0.096	0.077	1.000	0.131	0.063	.094
DI	0.103	0.080	0.092	0.063	0.131	1.000	0.120	.098
DD	0.046	0.068	0.052	0.033	0.063	0.120	1.000	.064

TABLE 6
Asymmetric Pairwise Overlaps*

	AA	TT	TA	ST	II	DI	DD	AVG.
Version - Most Relevant								
AA	1.000	0.329	0.401	0.496	0.340	0.368	0.266	0.367
TT	0.286	1.000	0.328	0.293	0.348	0.332	0.323	0.318
TA	0.451	0.424	1.000	0.520	0.355	0.420	0.344	0.419
ST	0.459	0.312	0.428	1.000	0.284	0.332	0.234	0.341
II	0.361	0.424	0.334	0.325	1.000	0.508	0.365	0.386
DI	0.346	0.359	0.351	0.337	0.450	1.000	0.490	0.389
DD	0.192	0.268	0.221	0.183	0.248	0.376	1.000	0.248
AVG	0.349	0.353	0.344	0.359	0.338	0.389	0.337	
Version - All relevant								
AA	1.000	0.276	0.348	0.381	0.275	0.323	0.233	0.306
TT	0.223	1.000	0.237	0.212	0.258	0.274	0.268	0.245
TA	0.361	0.304	1.000	0.402	0.281	0.310	0.241	0.316
ST	0.379	0.261	0.385	1.000	0.233	0.247	0.172	0.279
II	0.297	0.344	0.292	0.254	1.000	0.418	0.292	0.316
DI	0.305	0.319	0.283	0.235	0.366	1.000	0.458	0.328
DD	0.178	0.253	0.178	0.132	0.207	0.370	1.000	0.220
AVG	0.291	0.293	0.287	0.269	0.270	0.324	0.277	
Version - All Documents								
AA	1.000	0.145	0.250	0.229	0.210	0.193	0.103	0.188
TT	0.103	1.000	0.113	0.088	0.140	0.131	0.123	0.116
TA	0.265	0.169	1.000	0.262	0.188	0.180	0.119	0.197
ST	0.259	0.141	0.279	1.000	0.159	0.131	0.080	0.175
II	0.193	0.182	0.163	0.129	1.000	0.230	0.131	0.171
DI	0.180	0.172	0.158	0.108	0.233	1.000	0.240	0.182
DD	0.078	0.131	0.085	0.053	0.108	0.194	1.000	0.108
AVG	0.180	0.157	0.175	0.145	0.173	0.177	0.133	

*The representations in the columns form the denominator of the overlap measure.

TABLE 7
Union Pairwise Overlaps

	AA	TT	TA	ST	II	DI	DD	AVG.
Version - Most Relevant								
AA	0.328	0.520	0.549	0.481	0.558	0.523	0.502	0.495
TT	0.520	0.285	0.533	0.500	0.512	0.491	0.446	0.470
TA	0.549	0.533	0.369	0.525	0.594	0.548	0.525	0.519
ST	0.481	0.500	0.515	0.304	0.553	0.510	0.485	0.478
II	0.558	0.512	0.594	0.553	0.348	0.500	0.499	0.509
DI	0.523	0.491	0.548	0.510	0.500	0.309	0.430	0.473
DD	0.502	0.446	0.525	0.485	0.499	0.430	0.237	0.446
Version - All Relevant								
AA	0.283	0.449	0.475	0.457	0.505	0.465	0.449	0.441
TT	0.449	0.229	0.453	0.451	0.456	0.424	0.388	0.407
TA	0.475	0.453	0.294	0.462	0.514	0.479	0.458	0.448
ST	0.457	0.451	0.462	0.281	0.516	0.483	0.461	0.445
II	0.505	0.456	0.514	0.516	0.306	0.462	0.459	0.460
DI	0.465	0.424	0.479	0.483	0.462	0.268	0.385	0.424
DD	0.449	0.388	0.458	0.461	0.459	0.385	0.216	0.402
Version - All Documents								
AA	0.220	0.353	0.395	0.412	0.380	0.386	0.369	0.359
TT	0.353	0.156	0.363	0.384	0.331	0.335	0.302	0.318
TA	0.395	0.363	0.234	0.418	0.398	0.402	0.380	0.370
ST	0.412	0.384	0.418	0.249	0.420	0.428	0.402	0.388
II	0.380	0.331	0.398	0.420	0.203	0.361	0.347	0.349
DI	0.386	0.335	0.402	0.428	0.361	0.206	0.332	0.350
DD	0.369	0.302	0.380	0.402	0.347	0.332	0.166	0.329

is, if a retrieval environment were limited to a single representation, which one would it be? If a second could be added, which of the remaining six representations contribute the most over and above the effect of the first representation? A third representation could be added over and above the first two, and then a fourth representation, and so on.

The most sensible measure to use in answering this question is the union overlap. Tables 8 and 9 present the results of this analysis. Table 8 uses all seven representations and analyzes both the highly relevant as well as the total relevant measures across queries. Since three representations (TA, DI, ST) are composed of other representations, the analysis was repeated in Table 9 omitting these "compound" representations.

Tables 8 and 9 present four different models -- different orderings of representations. Such models, if consistent, would allow a searcher to know which combinations of fields would be most likely to retrieve relevant documents. Such models would also point to obvious economies in the design and operation of retrieval systems. Unfortunately, these data suggest that the models are not consistent. What appears to be highly consistent, however, is the cumulative increase in the percentage of relevant

TABLE 8
Representations Ordered by Incremental Improvement

Version - Most Relevant							
Order	1st	2nd	3rd	4th	5th	6th	7th
Representation	TA	II	AA	DD	TT	ST	DI
No of Documents	299	444	574	656	722	768	810
Cum. Percentage	.369	.548	.709	.810	.891	.948	1.000

Version - All Relevant							
Order	1st	2nd	3rd	4th	5th	6th	7th
Representation	II	ST	DI	TA	TT	AA	DD
No of Documents	527	889	1118	1318	1466	1602	1723
Cum. Percentage	.306	.516	.649	.765	.850	.930	1.00

TABLE 9
Representations Ordered by Incremental Improvement*

Version - Most Relevant				
Order	1st	2nd	3rd	4th
Representation	II	AA	TT	DD
No. of Documents	282	452	554	634
Cum. Percentage	.348	.558	.684	.783

Version - All Relevant				
Order	1st	2nd	3rd	4th
Representation	II	AA	DD	TT
No. of Documents	527	870	1093	1275
Cum. Percentage	.306	.505	.634	.740

*Compound representations omitted.

documents accounted for as each additional representation is included. This similarity may simply be due to the fact that the four models are based on highly interrelated data -- data that are subsets of one another. When the cumulative percentages are plotted against the order, the resulting curves appear to be Zipfian in form and when broken down according to Bradford's law of scatter, the obtained proportions are 1:3:7. The theoretical proportions could easily be in the form 1:3:9, but no attempt was made to verify this analytically.

An ancillary question is that of unique contribution of the different representations. That is, for a given representation, what documents does it contribute to the relevant retrieved that were not retrieved under any other representation? The question is equivalent to the observed improvements in the models when the representation is the last entered into the model. Tables 10 and 11 report incremental improvement for each representation, assuming the representation entered the model first or last. These are the maximum and minimum incremental improvements for each representation. Again, the index phase is distinctively unique, but more so under the full model than under the restricted one. Table 11 shows AA's unique contribution to be equivalent to II when the overlaps with the compound field (of which AA was a part) are not included in the model. These systematic differences in incremental improvement suggest that the patterns of overlap may be

TABLE 10
Recalls and Unique Contributions
of 7 Representations

Reps.	Entered 1st* No. of Docs	%	Entered Last* No. of Docs	%
Version - Most Relevant				
AA	266	.328	49	.060
DD	192	.237	44	.054
DI	250	.309	42	.052
II	282	.348	74	.091
ST	246	.303	44	.054
TA	299	.369	53	.065
TT	231	.285	52	.064
				.440
Version - All Relevant				
AA	488	.283	137	.080
DD	373	.216	127	.074
DI	462	.268	120	.070
II	527	.306	196	.114
ST	485	.281	149	.086
TA	506	.244	134	.078
TT	395	.229	133	.077
				.579

*Entered 1st is the equivalent of recall-1 across queries when no overlap is taken into account. Entered last are the unique documents found only by that representation.

TABLE 11
Unique Contributions of 4 Representations*

Rep.	No of Docs	%	No of Docs	%
	Version-Most Relevant		Version-All Relevant	
AA	125	.196	269	.210
DD	85	.133	197	.154
II	114	.178	271	.213
TT	88	.138	182	.143

*Recalls on 1st entered are same as in TABLE 10.
Compound representations excluded.

representation specific. It should be noted though, that the best unique contributor, II, in the full model retrieved only 20% (i.e. .091/.44) of the uniquely found documents and performed at the .35 recall level. Table 10 also reports the sum of the unique percentages, 44% for the rel-1 measure, 58% for rel-2. In other words only 56% and 42% of the documents were overlapped; another indication of the low probability of overlap observed in this and other studies.

Lastly, it is important to restate the difficulty of clearly interpreting the overlap measures. As previously mentioned, representations may be confounded with searchers.

VI. PHASE II PLANS

The second phase of the representation project is designed to 1) replicate the observations and findings of the first phase, 2) develop models that account for the results of the first phase and 3) test these in the experimental environment of the second phase. This section describes anticipated changes and extensions of the study methodology that will be incorporated in the second phase.

1. Data Base: The data base for the second phase will be a portion of the 1980 PsycInfo data base produced by the American Psychological Association: the printed counterpart is Psychological Abstracts. 12,000 records will again be used; dissertations will be excluded from the loaded data base. PsycInfo was selected as a "soft" data base with a different user population, in order to test the generalizability of the INSPEC study results. Additionally, PsycInfo records contain the same four fields that constituted the representations: descriptors, title, abstract and a free text index phrase. A user population for PsycInfo and searchers experienced with the data base are readily available. The DIATOM programs will again be used.

2. Research Design: The latin square design controlled for searcher differences on the performance dependent variables, but not on the overlaps. A different research design will

be used in order to obtain estimates of overlap attributable to (1) representations and (2) searchers.

In order to obtain searches on the same query, and the same representation for all searchers, the number of levels of representations and searchers probably will be reduced; the four primary representations will be maintained: title, abstract, index phrase and descriptors; four searchers will be used to obtain a balanced design.

3. Procedures: Procedures will parallel those of the first phase, revised to meet the requirements of the research design. This will be achieved by using some form of a completely crossed factorial design.

4. Models: A major activity of Phase II will be the development and analysis of models that account for the observed findings. Our current interest is in probabilistic models: by chance alone what is the minimum and maximum overlaps among representations that could be expected for a given data base. For the minimum overlaps we can proceed by assuming complete independence of representations and by using the relative frequency of each representation, we can determine the probability that random samples of two representations will contain documents in common.

The maximum overlaps can be calculated from an analysis of the number of unique words (types) in each representation. For example, in a sample of 1500 documents in the INSPEC data base, there are 9674 unique words in the abstracts (AA), but only 3481 types in the titles (TT). This lower number clearly puts an upper limit on the overlap between the two representations. Truncation must be excluded from consideration in this type of analysis; otherwise there will not be any real limit on the maximum possible overlap.

When this analysis is completed, other types of models need to be explored -- particularly models which will attempt to predict the performance-overlap results of both phases of this project.

5. Activity: The data in this report will continue to be analyzed by the project staff and consultants identified in the proposal. Data collection for hypothesis testing will go on as the second phase is implemented, (e.g. data base characteristics including distribution of terms in the representations, and distribution of search technique by representation and by searcher). Again, the emphasis will be on representations rather than searchers or searches; searcher difference will be incorporated only as necessary to control the variable in the overlap measures.

REFERENCES

AITCHISON, T.M.; HALL, A.M.; LAVELLE, K.J.; TRACY, J.M; Comparative Evaluation of Indexing Languages, Part II: Results, Project INSPEC, Institute of Electrical Engineers, London, England, 1970.

BATTEN, WILLIAM E. Document Description and Representation. In: Cuadra, Carlos, Annual Review of Information Science and Technology. Volume 8, American Society for Information Science, Washington, DC, 1973.

CLEVERDON, C.W. The Cranfield Tests on Index Language Devices. ASLIB Proceedings, 19, No. 6, June 1967. Pp. 173-194.

CLEVERDON, C.W.; MILLS, J.; KEEN, M. Factors Determining the Performance of Indexing Systems, 2 Volumes, College of Aeronautics, Cranfield, England, 1966.

HARRIS, JESSICA L. Document Description and Representation, In: Cuadra, Carlos, Annual Review of Information Science and Technology, Volume 9, American Society for Information Science, Washington, DC, 1974.

KEEN, E. MICHAEL. The Aberystwyth Index Language Test. The Journal of Documentation, Volume 29, No. 1, March 1973, pp. 1-35.

REFERENCES, Continued

KEEN, E. MICHAEL. Review of PRECIS, LCSH, and KWOC: A Report of a Research Project designed to Examine the Applicability of PRECIS to the Subject Catalog of an Academic Library. By Roslyn Hunt et al., Journal of Documentation, Vol. 34, No. 4, December 1978, pp. 356-357.

KOLL, M.; McGILL, M.J.; NOREAULT, T. Individual Differences in OnLine Searching. Paper presented at the 1979 ACM Computer Science Conference. February 1979, Dayton, Ohio.

McGILL, MICHAEL J. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. Research Proposal submitted to the National Science Foundation, January 1978.

McGILL, MICHAEL J. Knowledge and Information Spaces: Implications for Retrieval Systems. Journal of the American Society for Information Science, Vol. 27, No. 4, July/August 1976, pp. 205-210.

McGILL, MICHAEL J.; HUITFELDT, JENNIFER. Experimental Techniques in Information Retrieval. In: Williams, Martha, Annual Review of Information Science and Technology, Vol. 14, American Society for Information Science, Washington, DC, to be published 1979.

REFERENCES, Continued

NOREAULT, T.; McGILL, M.J.; KOLL, M. A Comparison of Manual Versus Automatic Indexing for Bibliographic Retrieval Systems. Paper presented at the 1979 ACM Computer Science Conference, February 1979, Dayton, Ohio.

RICHMOND, PHYLLIS. A Review of the Cranfield Project. American Documentation, Volume 14, No. 4, 1963, pp. 307-311.

SALTON, GERARD. A New Comparison Between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART). Journal of the American Society for Information Science, Vol. 23, No. 2, March-April 1972, pp. 75-84.

SMITH, LINDA C. Selected Artificial Intelligence Techniques in Information Retrieval Systems. Dissertation, School of Information Studies, Syracuse University, April 1979.

SPARCK JONES, K.; BATES, R.G. Research on Automatic Indexing. 1974-1976. Volume 1. Text, Computer Laboratory, University of Cambridge, Cambridge, England, 1977.

SWANSON, D.R. Searching Natural Language Text by Computer. Science Volume 132, No. 3434, October 21, 1960, pp. 1099-1104.

REFERENCES, Continued

TAUBE, MORTIMER. Evaluation of Information Systems for Report Utilization. In: Studies in Coordinate Indexing, Volume 1, Documentation, Inc. Washington, DC, 1953.

VICKERY, BRIAN C. Document Description and Representation. In: Cuadra, Carlos, Annual Review of Information Science and Technology, Volume 6, Encyclopedia Brittanica, Inc. Chicago, IL, 1971.

APPENDICES

Appendix A - Training Materials

- Project Description
- Searcher's Job
- Data Base
- DIALOG-Simulator Differences
- The Representations
- 003-Practise Search
- 004-Practise Search

Appendix B - Form for Relevance Judgments

Appendix C - Directions to Users

- NSF Information Retrieval Project, 2 pages
- Query Form

Appendix D - Forms for Searcher, Attached to Query, 2 pages

Appendix E - Latin Square Design, four pages

Appendix F - AOV Summary Results

- Recall-1
- Recall-2
- Precision-1
- Precision-2
- Tot-Ret.

Appendix A

PROJECT DESCRIPTION

This project will examine the relation between the relevance of retrieved citations and the fields that were searched to obtain them. Retrieval from seven different document representations will be studied. A representation consists of one or two designated search fields.

The data base for the study is Computer and Control Abstracts (a subfile of INSPEC). The system you will use is a local simulator of DIALOG, mounted on the S.U. computer. Almost all DIALOG features are available for you to use, but some restrictions will be made to achieve the study objectives.

The objectives of the study require you to conduct high recall searches, but with a limit of no more than 50 citations per query.

In all, you will be asked to search 98 queries. Over the course of the study, you will use all seven representations, but for each query only one representation will be assigned.

For each query, you will be asked to search from a request form; the statement of the query was prepared by a real user who will receive the output. The request form will also prescribe the representation you are to use. The unique password assigned to the request will automatically "lock" the search so that you can only search on the designated parts of the citations.

After you have completed each search (including the essential print command), return the search request form and a copy of your interaction with the system to Brian McLaughlin.

(5/2/80)

Appendix A

SEARCHER'S JOB

Your job as a searcher on this project will be to prepare and carry-out a high recall search for each request using one of the seven representations as specified.

You will receive the query statement as it was written by the requestor. This will be the only information you will receive regarding the user's request since there will be no face-to-face or telephone negotiations between you and the user.

One of the seven representations will be designated on the request form. The computer will be restricted to conduct the search using that representation, therefore your search strategy should be planned accordingly. You will be given a thesaurus for controlled vocabulary descriptor searching.

You may perform the search on any terminal that is or can be connected to Syracuse University, that is convenient for you, as long as hard copy can be printed. You are to perform a high-recall search with fifty citations as a maximum. You will be expected to complete the search within 48 hours after receiving the request form. Then return (1) the search request form - filling in the needed information, and (2) a copy of your interaction with the system.

NOTE: Limit the use of the thesaurus to this study only. We are legally bound by our contract to this limitation.

(5/2/80)

DATA BASE

Computers and Control Abstracts is that portion of the INSPEC Data Base dealing with all areas of computing and information science. The specific data base that will be searched in this study consists of four months (Sept. - Dec. 1979) of Computer and Control Abstracts.

The citations you will retreive will be organized as follows:

DNnumber (abstract numbers from INSPEC journals)
Title
Authors (separated by commas)
Source field: as follows
Publication: (volume and issue number)(part number) pagination data
Following this may be information in [ ]. This is information on the cover-to-cover translation as follows: [publication; (volume and issue) pages date] (type of unconventional media)(availability) (Title of conference), (location of conference); (sponsoring organization) (date) language
Abstract
Indexing information

NOT all the citations will contain each of these items of information.

DIALOG - SIMULATOR DIFFERENCES

The DIALOG simulator you will be using to conduct the searches is almost identical to "regular" DIALOG. In general, searching should be performed in the same way as any DIALOG search.

The restrictions, cautions and limitations are noted below.

1. Each new query you search must be started with the full BEGIN.

2. To restrict a search to a particular language, use a Limit /ENG (for English), or whatever language you wish.

3. Adjacency (nW) cannot be used with either truncation or stemming.

4. Adjacency may run very slow; the field operator (F) can be used instead.

(5/2/80)

THE REPRESENTATIONS

Appendix A

You will be using seven different representations during the study. A representation names the one or two fields of the citation to which your search must be restricted. You will search on only one representation for any given query. The representation you are supposed to search on will be designated on the request form we give to you. A unique password will be given with each request and this password will automatically lock the search onto the assigned representation.

The seven representations and the fields they will search are as follows:

TT - will search terms in title only.

AA - will search terms in abstract only.

DD - will search descriptor terms only. A thesaurus will be provided to you for use with this controlled vocabulary representation. (The thesaurus may only be used on this project).

II - will search identifier terms only.

TA - will search terms in title and abstract only.

ST - will search stemmed terms in title and abstract only. The computer will automatically take the logical root of any entered term. Truncation cannot be used with this representation.

DI - will search terms in descriptor and identifier fields. The thesaurus will be provided for use with this controlled vocabulary representation.

One representation with which you may be unfamiliar is stemming (ST), which will be used with title and abstract words only. A stemmed term is a word that has been shortened by the computer to its logical root. This is similar to truncation in that the stem LIBRAR would retrieve LIBRARY, LIBRARIES, LIBRARIAN, etc. For truncation however, the root is determined by the searcher. For example, if you entered LIBRARY under the ST representation, the computer would automatically be reduced to its logical root and LIBRARY, LIBRARIES, LIBRARIAN, LIBRARIANS, etc. would all be retrieved.

Truncation is not to be used with the stemming representation. In fact, the simulator will reject any attempts to use truncation in this representation.

(5/2/80)

Query # 003 - Practice Search

Appendix A

NAME: _____ DATE: _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;

I am interested in information about voice recognition systems and the used of speech recognition in man-machine systems. I am particularly interested in the use of interactive terminals and continuous speech recognition. I do not want citations that deal only with computer pattern recognition. The information must also include voice recognition.

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL. 4/4/80

Query # 004 - Practice Search

Appendix A

NAME: _____ DATE: _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;

My topic of interest involves national and international policy issues as they relate to computers and information. I would like information about ~~the~~ how the political structure affects the communications market and how different policies affect database usage, applications, and cost. Although I am especially interested in policies with regard to management information systems and EDP management, I would like as many citations as possible concerning the broader area of policy issues.

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL. 4/4/80

NSF INFORMATION RETRIEVAL PROJECT

INSTRUCTIONS TO PARTICIPANTS

Attached you will find a copy of your interest statement and two copies of a list of references. List (a) is to be used as part of the study and should be returned after you make your judgements of relevance. Copy (b) is yours to keep.

Each citation is organized into seven parts:

DN - Document identification number
TI - Title
AU - Author
SO - Source of the citation (i.e. journal title)
AB - Abstract
DT - Date
DE - Descriptors of the citation

Please read each citation and abstract to form an idea of what that particular document (book, article, report) is about. Compare this to your interest statement, and for each citation listed, decide how closely that citation is related to your topic. Based on the information in front of you, is the citation relevant to your topic, or not relevant to what you had in mind.

Use the following scale for your judgement:

1 - Definitely relevant to your topic.

2 - Probably relevant to your topic.

3 - Probably not relevant to your topic.

4 - Definitely not relevant to your topic.

Please rate each citation by placing the number corresponding to your judgement in the box immediately following each citation. After you have checked all the citations to see whether or not they are relevant to your interest statement, please return the copy with the judgements to us in the pre-addressed envelope through campus mail. If you are not on campus, these envelopes should be used to return the completed forms to us through the regular mail service. Thank you for your cooperation.

If you have any questions, please contact us at:

School of Information Studies
Syracuse University
113 Euclid Avenue
Syracuse, New York 13210
423-~~4522~~ 4549

6/16/80

SYRACUSE UNIVERSITY
SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

NSF INFORMATION RETRIEVAL PROJECT

We are working on a project which will help us understand how the pertinence of information retrieved by computer is related to the method by which it is searched.

For this project, we need information requests which will be searched in Computer and Computer Control Abstracts (from October 1979 to January 1980). If you need information in the area of computers and information science, we will conduct a search for you free of charge. All you have to do is submit a search request to us and give us information on how we did after the search.

For the search request we would like you to describe a topic of interest to you; one you are working on or are familiar with, in the computer field. Several days later you will receive a list of citations that have been retrieved by computer. You will be asked at that time to indicate which of these are pertinent to your interest. One copy of the computer output will be returned to us and the other copy will be for your own use.

We would very much appreciate your cooperation and participation in this project. If you are willing to participate, please read the attached pages and write your search request in the space provided.

If you do not need a search, please pass this form to a student.

7/24/80

SYRACUSE UNIVERSITY

SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

NSF INFORMATION RETRIEVAL PROJECT

As a participant in this project we would like you to submit a search request (on the attached form) about some aspect of computers and information science.

We will take your request and search the current issues of COMPUTER AND COMPUTER CONTROL ABSTRACTS. The results of this search will be a list of citations to books and journal articles.

We will then give you this list of citations and ask that you let us know which of these are most pertinent to your search request.

* * * * * * * * * * * * *

The enclosed form is for you to describe your topic of interest. If you are planning a talk or doing a paper, you probably have a topic in mind; if you don't have a topic you are working on, consider one with which you are familiar. Using this form, write down your information requirements as if you were talking to a colleague who understands the field as well as you do. Don't worry about trying to say it in "computerese"; we have trained people to make sure that your search is conducted professionally.

* * * * * * * * * * * * *

Thank you for your cooperation. If you have any questions, please feel free to contact us.

NSF Information Retrieval Project
School of Information Studies
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-4522

4/4/80

NAME:_____ DATE:_____

SCHOOL ADDRESS:_____ PHONE:_____

HOME ADDRESS:_____ PHONE:_____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Given your purposes in requesting this search, how many citations do you want?_____

About how many citations on your topic do you expect to receive from this computer search?_____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL. 4/4/80

SEARCH QUERY COVER SHEET

Searcher:_____ Search Query Number_____

Date to Searcher:_____ Representation Code this Query:___

Date to be Returned:_____ DIALOG Password_____

Some Important Notes:

1. Each new query to be searched must be started by the full BEGIN command.

2. You do not need to LOGOFF after each query before starting the next query. You do need to PRINT the documents retrieved before typing the BEGIN command for the new query.

3. Truncation cannot be used with the stemming representation (ST); it can be used with other representations.

4. Though you can use adjacency, you should know that it may run very slowly. Instead, you may choose to use the field operator (F). This implementation of DIALOG will not allow the use of adjacency with truncation, or adjacency with stemming.

To LOGON and LOGOFF

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the computer is given below.

Everything you type at the terminal must be sent to the computer with a carriage return.

The computer responses to some of these commands are not given here.

---

1. If you are using a dial-up terminal, the phone number is 423-1313. Remember, it must be a hard-copy terminal.

2. Turn power on and hit carriage return.

3. Type: LOG 3434,14

4. Type: NSF

5. Type: DO DIALOG

   The computer will ask for your dialog password. It is given at the top of this page.

Date Returned to Brian McLaughlin:_____ Date Returned to NSF:_____

(5/2/80)

SEARCH QUERY COVER SHEET - Page 2

6. Type: BEGIN

   The computer will ask for the query number and the representation code. Both can be found at the top of Page 1.

7. Carry out the search for this query.

   Remember, we want a high recall search with a maximum of 50 documents retrieved.

   Before starting a new query you need to have the set of retrieved documents printed. Use the PRINT command; the format should always be 1.

8. If you want to search another query, look at the COVER SHEET for that query and begin at Step 6.

   If you are completely done searching for now, go to Step 9.

9. Type: LOGOFF

10. Type: K/F

11. Turn power off, collect your materials and submit them to Brian McLaughlin.

Submitting Searches

Brian McLaughlin will distribute and collect all searches. When a search is completed, you need to submit this COVER SHEET and a copy of your interaction. Queries should be searched and returned within 48 hours after receiving them.

Help and Assistance

1. Brian McLaughlin
   210 Hubbell Avenue
   Syracuse, New York
   476-7359 (Home)
   423-2091 (Work)

2. NSF Retrieval Project
   113 Euclid Avenue
   Syracuse, New York
   423-4522

(5/2/80)

14 LS T

SQUARE 1

	101	102	103	104	105	106	107
EDWA	DD	AA	TA	DI	ST	TT	II
VAUG	ST	II	AA	DD	TT	TA	DI
MINO	DI	TA	TT	II	DD	ST	AA
SETT	TA	DD	DI	TT	AA	II	ST
LAUB	AA	ST	DD	TA	II	DI	TT
MCLA	II	TT	ST	AA	DI	DD	TA
ABBO	TT	DI	II	ST	TA	AA	DD

SQUARE 2

	108	109	110	111	112	113	114
EDWA	II	DD	ST	DI	AA	TA	TT
VAUG	AA	DI	DD	II	TA	TT	ST
MINO	DI	ST	TT	DD	II	AA	TA
SETT	DD	TT	TA	ST	DI	II	AA
LAUB	TT	AA	II	TA	ST	DD	DI
MCLA	ST	TA	AA	TT	DD	DI	II
ABBO	TA	II	DI	AA	TT	ST	DD

SQUARE 3

	115	116	117	118	119	120	121
EDWA	DD	ST	DI	AA	TT	II	TA
VAUG	AA	II	TA	ST	DI	TT	DD
MINO	ST	TT	DD	II	TA	DI	AA
SETT	TT	TA	ST	DI	AA	DD	II
LAUB	TA	AA	TT	DD	II	ST	DI
MCLA	II	DI	AA	TT	DD	TA	ST
ABBO	DI	DD	II	TA	ST	AA	TT

SQUARE 4

	122	123	124	125	126	127	128
EDWA	TA	ST	II	TT	DI	AA	DD
VAUG	DD	II	TT	DI	TA	ST	AA
MINO	DI	AA	ST	II	TT	DD	TA
SETT	AA	TT	DI	TA	DD	II	ST
LAUB	II	TA	DD	AA	ST	DI	TT
MCLA	TT	DD	AA	ST	II	TA	DI
ABBO	ST	DI	TA	DD	AA	TT	II

SQUARE 5

	129	130	131	132	133	134	135
EDWA	DI	II	TA	DD	AA	TT	ST
VAUG	TT	ST	DI	TA	DD	II	AA
MINO	II	AA	TT	DI	TA	ST	DD
SETT	ST	DD	II	TT	DI	AA	TA
LAUB	TA	TT	DD	AA	ST	DI	II
MCLA	DD	DI	AA	ST	II	TA	TT
ABBO	AA	TA	ST	II	TT	DD	DI

SQUARE 6

	136	137	138	139	140	141	142
EDWA	TT	TA	ST	DI	II	AA	DD
VAUG	ST	TT	DD	II	AA	TA	DI
MINO	AA	II	TA	ST	DD	DI	TT
SETT	TA	AA	TT	DD	DI	II	ST
LAUB	DI	DD	II	TA	TT	ST	AA
MCLA	DD	ST	DI	AA	TA	TT	II
ABBO	II	DI	AA	TT	ST	DD	TA

SQUARE 7

	143	144	145	146	147	148	149
EDWA	TA	TT	ST	II	DI	AA	DD
VAUG	DD	DI	II	TT	TA	ST	AA
MINO	DI	II	AA	ST	TT	DD	TA
SETT	AA	TA	TT	DI	DD	II	ST
LAUB	II	AA	TA	DD	ST	DI	TT
MCLA	ST	DD	DI	TA	AA	TT	II
ABBO	TT	ST	DD	AA	II	TA	DI

SQUARE 8

	150	151	152	153	154	155	156
EDWA	II	TT	DD	AA	TA	DI	ST
VAUG	DD	AA	TT	DI	II	ST	TA
MINO	TA	DD	II	TT	ST	AA	DI
SETT	ST	II	TA	DD	DI	TT	AA
LAUB	DI	TA	ST	II	AA	DD	TT
MCLA	AA	ST	DI	TA	TT	II	DD
ABBO	TT	DI	AA	ST	DD	TA	II

SQUARE 9

	157	158	159	160	161	162	163
EDWA	AA	ST	II	DI	TA	TT	DD
VAUG	TT	DI	TA	AA	ST	DD	II
MINO	ST	II	TT	TA	DD	DI	AA
SETT	II	TT	DI	DD	AA	TA	ST
LAUB	DD	AA	ST	TT	DI	II	TA
MCLA	DI	TA	DD	ST	II	AA	TT
ABBO	TA	DD	AA	II	TT	ST	DI

SQUARE 10

	164	165	166	167	168	169	170
EDWA	AA	TT	DI	ST	DD	II	TA
VAUG	DI	AA	ST	TA	TT	DD	II
MINO	TT	DD	AA	DI	II	TA	ST
SETT	ST	DI	TA	II	AA	TT	DD
LAUB	DD	II	TT	AA	TA	ST	DI
MCLA	TA	ST	II	DD	DI	AA	TT
ABBO	II	TA	DD	TT	ST	DI	AA

SQUARE 11

	171	172	173	174	175	176	177
EDWA	TT	ST	DI	II	AA	TA	DD
VAUG	ST	DD	II	AA	TA	TT	DI
MINO	II	AA	TT	ST	DD	DI	TA
SETT	AA	TA	ST	DD	DI	II	TT
LAUB	DD	DI	AA	TA	TT	ST	II
MCLA	TA	TT	DD	DI	II	AA	ST
ABBO	DI	II	TA	TT	ST	DD	AA

SQUARE 12

	178	179	180	181	182	183	184
EDWA	AA	TT	TA	DI	DD	II	ST
VAUG	DI	AA	II	ST	TT	DD	TA
MINO	TT	DD	ST	AA	II	TA	DI
SETT	DD	II	DI	TT	TA	ST	AA
LAUB	II	TA	AA	DD	ST	DI	TT
MCLA	ST	DI	DD	TA	AA	TT	II
ABBO	TA	ST	TT	II	DI	AA	DD

SQUARE 13

	185	186	187	188	189	190	191
EDWA	TA	II	TT	AA	ST	DI	DD
VAUG	DD	TT	DI	ST	II	TA	AA
MINO	AA	DI	TA	II	TT	DD	ST
SETT	ST	TA	DD	TT	DI	AA	II
LAUB	II	DD	AA	DI	TA	ST	TT
MCLA	DI	ST	II	DD	AA	TT	TA
ABBO	TT	AA	ST	TA	DD	II	DI

SQUARE 14

	192	193	194	195	196	197	198
EDWA	TT	DD	AA	DI	ST	TA	II
VAUG	DD	II	TT	AA	DI	ST	TA
MINO	DI	AA	ST	TA	II	DD	TT
SETT	II	TA	DD	TT	AA	DI	ST
LAUB	AA	TT	DI	ST	TA	II	DD
MCLA	ST	DI	TA	II	DD	TT	AA
ABBO	TA	ST	II	DD	TT	AA	DI

AOV SUMMARY TABLE: Recall-1

Source	Sum of Squares	df	Mean Square	F
Between Squares	2.624	11	.239	
Queries in Squares	10.415	58	.180	
Searchers	4.072	6	.679	
Squares X Searcher	7.940	66	.120	
Representations	1.415	6	.236	3.324*
Square X Representation	6.021	66	.091	1.282**
Residual (by subtraction)	19.714	276	.071	
Total	52.201	489		

*Region of rejection begins at 2.14 (α=.05) or 2.89 (α=.01)

**Region of rejection begins at 1.12 (α=.25). Since obtained value falls within the region of rejection, the square X representation source of variation is not pooled into the residual.

NOTE 1: Tukey's HSD region of rejection = 4.17
standard error - .0318

NOTE 2: Missing values in the data (14 queries retrieved no highly relevant documents) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

AOV SUMMARY TABLE: Recall-2

Source	Sum of Squares	df	Mean Square	F
Squares	.963	11	.088	
Queries in Squares	5.678	65	.087	
Searchers	4.088	6	.681	
Squares X Searchers	4.842	66	.073	
Representations	1.032	6	.172	3.44*
Pooled Error (by subtraction)	19.038	384	.050	
Total	35.641	538		

*Region of rejection begins at 2.14 (α=.05) or 2.89 (α=.01)

NOTE 1: Tukey's HSD region of rejection = 4.17
standard error = .0255

NOTE 2: Missing values in the data (7 queries retrieved no relevant documents at all) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

AOV SUMMARY TABLE: Precision-1

Sources	SS	df	MS	F
Squares	3.536	11	.321	
Queries in Squares*	15.066	72	.209	
Searchers	0.528	6	.088	
Squares by Searchers	3.740	66	.057	
Representations	0.219	6	.0365	.829 (n.s.)
Pooled error (by subtraction)	15.829	360	.044	
Total		521		

*Missing values in the data (66 cases with no documents retrieved) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed which results in more than one value for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum of squares. The approach is conservative in the sense that if the effect of representations were to be significant, it would also be significant if the other value for the Queries in Squares sum of squares were used.

AOV SUMMARY TABLE: Precision-2

Sources	SS	df	MS	F
Squares	5.489	11	.499	
Queries in Squares*	19.886	72	.276	
Searchers	0.691	6	.115	
Squares by Searchers	5.348	66	.081	
Representation	0.364	6	.0607	1.05 (n.s.)
Pooled Error (by subtraction)	20.788	360	.0577	
Total		521		

*Missing values in the data (66 cases with no documents retrieved) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed which resulted in more than one value for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum of squares. The approach is conservative in the sense that if the effect of representations were to be significant, it would also be significant if the other value for the Queries in Squares sum of squares were used.

Appendix F

AOV SUMMARY TABLE: Tot-Ret.

Sources	Sums of Squares	df	Mean Square	F
Between Squares	10688.347	11	971.668	
Queries in Squares	40273.878	72	559.359	
Searchers	19316.177	6	3219.363	
Squares X Searchers	13719.415	66	270.870	
Representations	3654.511	6	609.085	4.24*
Residual	61236.183	426	143.747	
Total	148888.51	587		

*Region of rejection begins at 2.14 (α =.05) or 2.89 (α =.01)

NOTE: Tukey's HSD region of rejection = 4.17;
standard error = 1.308